【开放学习前沿】

在线测评：过去，现在和未来

[英]萨莉·乔丹（Sally Jordan）

（英国开放大学物理系，英国）

【摘　要】文章阐述了广义上使用计算机的所有测评，重点说明了机考方式。测评任务的多样化、对即时反馈的诉求、阅卷的客观性以及节省资源的需要是撰写文章的驱动力。机考从早期单纯的多项选择题型和机读题型开始发展到了融合互操作性的便于学生在家里使用的网络系统形式。这种设计精细的在线测评系统由大学、公司设计开发并作为虚拟学习环境的一部分使用。例如，选择题可以通过某些技巧，如基于信任度的评分来减少其不足。在课堂上尤其是在同伴讨论中使用电子回复系统（应答器）很有效果。学生们设计的问题也会鼓励同伴之间围绕学习进行对话。

日趋精细的机考系统能够将数学题拆分成几步，并提供有针对性的、即时的反馈。文章也讨论了计算机代数以及简答配对题系统。计算机适性测验采用学生对以前问题的回答来改变测验的后续形式。更广泛地讲，在线测评包含了同伴测试、测评的电子文件夹、博客、维基和论坛。文章最后预测了在线测评的的未来发展：在线测评可在MOOCs（大规模开放远程课程）中使用；可用于学习分析；在线测评使教学、测评与学习逐渐融合；解放了人力测评，并能更加真实地测评考试内容。

【关键词】在线测评；机考；综述

【中图分类号】G720　**【文献标识码】**A　**【文章编号】**1008–7648（2015）05–0008–011

一、概述

广义上的在线测评（JISC 2006）是指任何有计算机参与的测评，可以是终结性、形成性或诊断性测评。因此，其范畴可包括网上提交的由辅导教师批改的作业，电子文件夹的测评，反思性博客，录制成声音文件的教师反馈，最常见的是机评测验。其他同类含义的表述包括技术辅助的测评，计算机支持的或计算机协助的测评。

在线测评的近期文献综述（Conole & Warburton, 2005; Dikli，2006; Hepplestone 等，2011; JISC, 2009; Kay & LeSage, 2009; Nicol, 2008; Ridgway 等,2004; Ripley ,2007; Stödberg, 2012）主要聚焦于技术辅助的测评或反馈。尽管本文侧重于机考，但仍采用广义上的定义。因在线测评这个概念非常宽泛，且因篇幅限制，不能谈及所有的在线测评系统，因此本文精选了几项技术进行介绍（例如电子投票系统或“应答器”）。本文也讨论了在线测评和纸质试卷的选择，重点介绍了物理学和相关学科领域中开发、使用或评估过的在线测评系统。

二、动力及发展

“有效的测评和反馈是指辅助学生在所选择的复杂学科领域里取得优异成绩的一种实践做法。学生参与有效的测评和反馈，这不但不会增加辅导教师的负担，还能帮助作为终身学习者的他们掌握学习技能，能够自信地继续学习。在这些目标的实现过程中，技术提供了巨大的支持。”（JISC2010，p.8）

在线测评是包含教学和测评方法的在线学习（Ashton&Thomas 2006, Gipps 2005）的自然组成

部分之一（Mackenzie 2009）。在线测评在设计上不断多样化并保证其真实性，比如，采用电子文件夹，模拟和互动游戏等方式，具有其他手段很难实现的测评技能（JISC 2010）。

学生无论何时何地选择做多项选择题，即便非常简单，也能检测出他们对题目所涉及的广泛内容的理解程度（Bull & Danson&McKenna 2004）。因此，学生只要有机会就可反复练习（Bull&Danson 2004），哪怕有时仅仅是练习同类题型的变形题（Jordan 2011）。在线测评能给出及时的反馈，并参照相关的课程内容对错误给出有针对性的解释（Jordan & Butcher 2010）。这成为学生、特别是远程学习的学生身边的‘“虚拟辅导教师’”（Ross 等人 2006）。在线测评允许学生私下犯错（Miller 2008），反馈是客观的（Miller 2008），不带有任何批判色彩（Beevers 等人 2010）。

研究表明定期的在线测试会提升年度考试成绩（Angus&Watson 2009）。在线测评增加了学生对课程的参与度，激发学习动力并帮助学生规划学习进度（Crebenik & Rust 2002, Jordan 2011）。学生可以根据网上作业检查对所学知识的理解从而制定未来学习计划，但是研究表明单纯的考试行为，即便是没有任何反馈的考试，也比多余的学习资料更能提升后续的学习表现。这就是所谓的测试效应，有关该领域的研究可参阅Roediger & Karpicke（2006）的文献。

所以说，在线测评对提高学生的学习具有很大的帮助。然而，有趣的是，“客观题”这个词，尤其是用来描述多项选择题，反应了早期采用多项选择题的目的是希望使测评更客观。最早期的多项选择题可能是e.l.桑代克的α和β测试，是第一次世界大战中美军用来测评新兵的服役情况（Mathews 2006）的。然而，20世纪，研究人员逐渐意识到写短文作为测试方法的局限性以后，多项选择题作为一种教学测验手段逐渐受到追捧（E.R.Bacon 2003）。Ashburn（1938）注意到不同教师对短文的评分不同，这种情形让人担心，类似的发现有很多（例如Millar 2005）。人工评分具有天生的不连续性，可能会因评分人对某个学生的期望而受到影响（Orrell 2008）。多项选择题具有客观性，而机评则具有连续性，这是多个评分者之间、甚或是一个评分者在不同时段评分都难以确保的特点（Bull & McKenna 2004, Butcher & Jordan 2010）。

尽管设计出高质量的问题也会被当作重要的任务（Bull & McKenna 2000），但是，机考在提高可信度的同时，既能节省时间又能节省资源（Dermo 2007）。机评尤其适合学生人数较多的大班（Whitelock & Brasher 2006），能让使用者更加充分地利用时间，从这一点来说，机评具有附加价值（JISC 2010）。

20世纪，大规模的多项选择题通过机读形式管理（如图1所示），学生在答题纸上标出每题所选答案。该系统（目前仍在使用）保证了测评的客观性，节省了资源，但是反馈的及时性和学生的参与度没有显现出来。根据Brown等人（1999）对高等教育实践做法的回顾，发现在线测评虽然也存在其他形式，但基本上都是“多项选择题”的同义词，设计者们仅仅是将测评从纸质版转移到了屏幕上，这已被证明是不恰当的做法。

图1 机读学生答题卡样式

大约从1980年开始，在线测评系统的使用呈快速增长的趋势，系统设计也日益精细。例如，由德比大学设计开发的三重测试系统（TRIADS）（Mackenzie 1999, TRIADS 2013）从1992年起连续使用至今。TRIADS特意包含了不同的题目类型从而辅助高阶技能的测试。

模块物理学教学软件（STOMP）测评系统的设计开发始自1995年（R.A.Bacon 2003）。最近的版本（Bacon 2011）是直接安装启用的QTIv2.1版，问题和测试的互操作性(QTI)版设计的目的是

交换编写工具、试题库和测评系统之间的试题、测试和结果的数据（IMS Global Learning Consortium 2013）。

20世纪90年代，物理学和工程学本科阶段学生的数学准备越来越受到关注，因此，催生了诊断性测试（Appleby 等人 1997, Appleby 2007）。诊断性测试采用的是技能层次法，基于学生对前一个问题的回答，利用专业系统来确定下一个问题该如何设计。因此，诊断性测试也是早期适应性测验的一个案例（详见第三章第9节）。

大约在本世纪初，考试开始转向使用在线测评系统，利用互联网将考试送达边远地区。英国开放大学（后文简称“英开”）于1997年利用Open Mark系统（详见第三章第6节）作为先驱，开发设计了一套互动问题，这是首次让学生尝试在线测评系统，而之前类似的问题都是通过CD-ROM发送给学生。直到2002年，大学才有足够的信心让在家学习的学生通过互联网的稳定支撑，使用网上问题进行在线测评（Jordan 等， 2003； Ross 等，2006）。学生的答案和分数都被记录在位于弥尔顿凯恩斯的大学总部服务器上。稳定的互联网系统也能使辅导教师评分的考试以电子方式上传和反馈（Freak 2008, Jordan 2011），这就消除了邮递系统所造成的延迟。

在商业领域，Questionmark（以前叫做“Question Mark Computing”）公司于1988年成立。Question Mark 网络版（1995年启动实施）被看作是世界上首个商业互联网测试产品。 其专业版于1993年启动实施，后来逐步被其认知版所取代（Kleeman 2013）。

随着越来越多的学习工具上线使用，大学和其他教育机构开始使用虚拟学习环境系统（VLEs），也通常被叫做学习管理系统。大部分的VLEs都融合了测评系统。例如，Moodle学习管理系统（Moodle 2013）于2002年首次上线使用，它的测验系统紧接着在2013年上线（Hunt 2012）。Moodle和它的测评系统（详见第三章第6节）都是开放资源，在影响在线测评工具发展的理念上反映了重大的变化。

三、当前机考发展状况

1. 选择题还是建构题

Hunt（2012）确认了Moodle 平台中使用的30多种不同的题型，例如拖拽题、计算题、数字题、判断正误题等题型更加多样化，但是Hunt从2 500多个Moodle 站点收集了5 000多万份专门的调查问卷显示：正在使用的大约90%的题型都是选择题，例如，多选题或给出备选答案供学生选择的拖拽题，远远超过需要学生自己组织回答的构建题。

关于选择题和构建题的优劣性，很多文献提供的证据存在明显的矛盾。选择题能够考察更宽泛的知识内容（Betts 等，2009； Ferrao， 2010），而建构题在其考察范围内可能更具有选择性。选择题的使用也避免了数据输入的问题，尤其是需要符号表示法的建构题中所遇到的问题，比如数学建构题就存在这样的问题（Beevers & Paterson， 2003；Jordan 等， 2003； Ross 等， 2006；Sangwin ，2013）。另外，选择题避免了配对题的不完整性和不准确性。偶尔也有建构类题没有被正确打分的情况（Butcher & Jordan 2010），Gill & Greenhow（2008）的研究报告指出了一个令人担忧的发现：学生们学会了在建构题答案中省略某些通常不被需要的或无法被测评系统识别的知识点，进而会在之后的答案中继续省略。

Conole & Warburton（2005）指出尽管有学者一直努力尝试（例如：Gwinnett & Cassella ，2011）使用选择题测评高阶学习成果，但这个方法具有一定的难度，而且，在一些多项选择题中，可以通过排除法来选择正确答案。这样，问题本身就无法如预想中一样去考核学习成果。例如，一道要求集成一个函数的问题，学生就可以通过区分所提供的备选来选择正确的答案（Sangwin， 2013）。对于所有的选择题来说，尤其是需要计算或代数处理的选择题，如果学生算出的结果在备选答案中找不到，那么，他们就会提前得到暗示：很可能他们的结果出现了错误（Bridgeman 1992）。甚至当检测一道设计得很好的有关力学概念的题目时（最早出现在Hestenes 等人1992的报告中），Rebello & Zollman (2004) 发现，同样的题目即便在开放式测验中，学生们答题的结果也与选择题中的任何一个备选答案不同。

在选择题中，学生们可以猜测答案，因此，老师们无从知晓学生们到底学习到了什么（Crisp, 2007）。 Downing （2003）并不关注对猜题获得的分数，他指出，单纯地通过猜题而通过整个考试是非常困难的。然而，Burton（2005）认为成功的猜题会对处于边缘学生的学习成果产生很大的影响。

Funk & Dickson （2011）使用完全相同的题目设计了多项选择题和简答题两种版本从而进行对比研究。50个学生尝试着对两个版本的每个题目进行回答，其中一半的学生先完成了10道简答题作为50道多选题的预备考试；而另一半的学生则先回答了50道多选题，作为多选题的后续又完成了10道简答题。在每个案例中，学生们在多项选择题中的表现远远好于回答简答题的表现（$p<0.001$）。然而，Ferrao（2010）发现，在多项选择题和开放式问答测验的分数之间有很高的关联度。一些学者称选择题特别适合特定学生群体，尤其是具有更多考试技巧的学生或更愿意冒险的学生（Hoffman 1967）。对两种题型的喜好也存在性别差异，例如，Gipps & Murphy（1994）发现在15岁年龄段的学生中，女孩不喜欢多项选择题，而男孩则喜欢选择题胜过简答题。Kuechler & Simkin （2003）发现母语为非英语的学生做多项选择题时在剖析词汇的细微差别方面会遇到困难。Jordan & Mitchell（2009）和Nicol（2007）明确了在选择题和建构题的回答中存在根本不同的认知过程。

对选择题最大的争议也许是来自于对其真实性的质疑。在评论医学院校广泛使用的多项选择题时，Mitchell 等人 （2003）曾引用Veloski（1999）的观点说：“不需要给患者介绍五种选择”。Bridgeman （1992， p. 271）在谈论工程师和化学师时也指出了同样的观点：他们很少“面对五个数字备选答案，然后选出唯一正确的解决方案”。

再回到1995年，Knight（1995, p. 13）指出“我们选择去测评什么以及采用什么样的方式来实现，彰显了什么是我们所珍视的”。Scouller（1998）指出使用选择题鼓励了学生采取浅层学习方法开展学习。尽管Kornell & Bjork（2007）没有证据表明学生认为简答题和短文作业题比多项选择题难，需要更加努力地理解这两种类型的题目。Roediger & Marsh（2005）和Marsh 等人 （2007）发现在使用多项选择题时有一种逐渐减弱的“测评效应”，这归因于学生在做多项选择题时记住的往往不是正确答案，而是错误选项。

对选择题有效性调查的结果存在明显的矛盾，这可能是因为所调查的问题不具有同质性（Simkin & Kuechler 2005）。不同的问题需要采用不同的类型来判断其有效性，有些问题（比如，选择三个相同的表述）则尤其适用于选择题。Burton （2005， p. 66）指出，很有可能具有特定格式和评分方式的考试，有时仅仅是因为其有瑕疵的选项和步骤而被判断为是不可靠的考试方式。不管采用什么类型的问题，重要的是要设计出高质量的问题（Bull & McKenna 2000）。例如，对于多项选择题来说，高质量就意味着除正确答案外的其他备选答案也要具有同等的貌似合理性。

即使是相对简单的多项选择题也能够制造“意外时刻”（Dermo & Carpenter 2011），Draper（2009）的催化测评观点就是建立在利用选择题刺激后续没有教师直接参与的更加深入的学习。有许多方式可以用来提高选择题的可靠性和有效性（Nicol 2007）。有关的方法将在第三章第2节至第4节详细讨论。

2. 基于信任度的评分和类似的方法

为了弥补学生们在多项选择题中猜测正确答案的事实，教师们采用了各种各样的方法。可采用单纯的否定打分（选择错误答案即减分）的方法，但是必须慎重（Betts 等， 2009； Burton ， 2005）。

Ventouras et al （2010）构建了一种为同一题目使用“成对的”多项选择题的考试类型（但这明显不是给学生使用的），该类考试的得分规则是，如果两个问题都答对了，将会获得“奖励”。考试的结果几乎与建构型题目考试的结果没有任何差别。McAllister & Guidice（2012）介绍了另外一种考试方法，就是将一系列的问题与备选答案相结合，这就会导致相对很长的答案列表（他们的案例是为50个问题设计了60个备选答案），这样就大大降低了猜测正确答案的可能性。然而，一般来说，为一系列的问题找到同样可行的备选答案是很困难的。

Bush（2001）介绍了一种“自由式多项选择题”的形式，在这种题型中，如果学生对某题的正确答案不确定，那么，他们可以选择不止一个答案来回答该问题。否定打分即是对错误答案扣分：每个正确答案是3分，每个错误答案被扣除1分，总分再除以3。如果一个学生知道某题的正确答案，那么他/她可以得分3/3，也就是说该题他/她可以得100%。如果一个学生认为某题的正确答案是两个选项之一，则该题可获得（3-1）/3，也就是说67%，而不是0%或100%。如果一个学生认为正确答案是三个选项之一，则该题可获得（3-2）/3，也就是33%，而不是100%或0%。毫无疑问，这种题型的打分方式是比较公平的，但是有些学生却认为这种方

式非常容易产生误解，而且会使学生们将重点放在猜题技巧上而非对正确答案知识的理解上。

长期以来，有一种观点认为考试分数的可信度可以通过适当地融入对学生的信任度而有所增加。Gardner-Medwin做了很多关于“以信任度为基础的”（或叫“以确信度为基础的”）评分，他指出这种方法不能支持一直可信任或一直不可信任的学生，却支持那些能够正确判断缘由的人。Gardner-Medwin使用了分数量和扣分原则（如表1所示）。

表1 以信任度为基础的评分和扣分原则

信任度	C=1（低）	C=2（中）	C=3（高）
如果正确应得分数	1	2	3
如果错误应扣分数	0	−1	−2

Rosewell（2011）要求学生在多项选择题给出之前指出自己的信任度指数，而Archer &Bates（2009）设计了信任度指数和自由文本框，要求学生给出所选答案的理由。Nix & Wyllie（2011）在形成性多项选择题中融合了信任度指数和反思博客，目的是要鼓励学生管理自己的学习经历。

3. 应答器

早在1970年以前，教室中就已经开始使用电子投票系统，也称作“观众反应系统”、“学生应答系统”和“应答器”。学生将选择题的答案输入到手持的“学生应答系统”，以此向教师传递想法，教师据此了解整个课堂学生的理解程度，从而适当地调整教学过程。Judson & Savada（2002）强调了前辈们如Boardman （1968）和Casanova（1971）早期的工作，他们都使用一种叫做instructoscope的硬线连接的系统。早期当Littauer（1972）注意到学生对课前提供问题的答案的辩论时，就确认了学生私下里输入反应的需求——这是此类方式的早期迹象，后来被Classtalk（Dufresne 等人 1996）和同伴指导法（Mazur 1991）所采用。有关应答器的使用有大量的文献综述，如Fies & Marshall (2006)，Caldwell (2007)，Simpson & Oliver (2007)和Kay & LeSage (2009)。网上教室如Blackboard Collaborate（Blackboard 2013），现在也能在虚拟环境中实现同样的应答反应功能。

许多作者（包括Wieman 2010）将学习的显著效果归因于应答器的使用，但是Fie & Marshall（2006）要求在这方面要做更加严格的研究，Beatty & Gerace（2009）认为应答器有不同的使用方法，而不应该将这些方法所带来的作用叠加到一起。研究发现同伴讨论是一种特别有效的方法，使课堂更具有互动性，学生们更加积极主动地参与到自己的学习过程中来（Dufresne 等，1996；Mazur， 1991； Crouch & Mazur， 2001；Lasry ，Mazur& Watkins ，2008）。“围绕学习进行的对话(Dialogue around learning)”是Nicol &Macfarlance-Dick（2006）所倡导的良好反馈实践七原则之一，Nicol（2007）建议这个原则可以通过启动多项选择题课堂讨论来实现。

4. PeerWise

Nicol（2007）指出“围绕学习的对话”可以通过小组协作、构建多项选择题的方式或通过评论其他人对考试的某些方面实现。PeerWise（Denny 等， 2008b；PeerWise， 2013）是由奥克兰大学计算科学系设计开发的系统，在该系统中，学生可创作自己的多项选择题，也可使用和评估同伴创作的题目。Luxton-Reilly & Denny （2010）描述了PeerWise背后的教学法，其前提是学生正从知识的消费者转向知识创造群和知识分享群的参与者。奥克兰的评估显示学生们对PeerWise系统的使用程度远远高于学校的要求（Denny 等， 2008c），学生们创作的问题的质量也非常高（Purchase 等，2010），在Peer Wise活动和后续撰写的问题(不仅是多项选择题)的表现之间存在很重要的关联性（Denny 等， 2008a）。爱丁堡大学物理航空学院一直重复使用这些成果（Bates 等， 2012；Bates & Galloway ，2013），Peer Wise活动和后续表现之间的关联被认为不仅适合于学习能力较弱的学生，也适合于学习能力较强的学生。

5. 计算机辅助的数学学习（CALM）和苏格兰测评项目（PASS-IT）：聚焦于将问题分解成“多个步骤（steps）”

海里特一瓦特大学从1985年开始实施计算机辅助的数学学习项目（CALM）（CALM 2001），各种不同的机考系统也部分源自于该项目，包括CUE、交互式历年论文（Interactive Past Papers）、苏格兰使用信息技术的考试项目（PASS-IT）、在线考试（i-assess）和NUMBAS（Foster 等，2012）。有些系统在具有高风险的终结性考试中使用，但是核心目的仍是支持学生的学习（Ashton 等，2006b）。从早期的观察来看，建构类题型受到了欢迎，因为此类题型可以给学生提供提示（Beevers & Paterson 2003）。

CALM考试系统的特征之一是使用“分解步骤”，即允许将问题分解为可管理的步骤，从而为不能继续考试的学生提供帮助（Beevers & Paterson，2003；Ashton 等，2006a）。图表2（a）中显示的是一个问题，一个学生选择没有任何中间步骤的帮助来回答问题，这样，在终结性考试中，他就可获得满分。同样地，如图2（b）所示，学生可点击“分解步骤”按钮，问题即可被分解成不同的步骤，这样，学生在终结性考试中通常能获得部分分数。

McGrire et al（2002）对比了学生采用CUE考试体系中三种不同格式（没有分解步骤，必须使用分解步骤，或者选择性地使用分解步骤）参加机考的结果和就同样内容采用纸质考试的结果。在这种情况下，采用分解步骤不存在扣分的情况。没有使用分解步骤的总分要低于使用分解步骤所获得的分数，而两者的分数均低于纸质考试的分数。他们的结论是，“这意味着没有分解步骤的情况下，纸质考试分数框架体系不能被当下所使用的机考框架体系所取代。题目越长越复杂，所产生的问题也越多。”他们没有在纸质考试所获得分数与不同格式的机考所获得分数之间发现任何不同。然而，他们认为，即使分数相同，“这也不意味着学生表现了同样的技能。另外，采用分解步骤为学生完成一道考题提供了指导。”

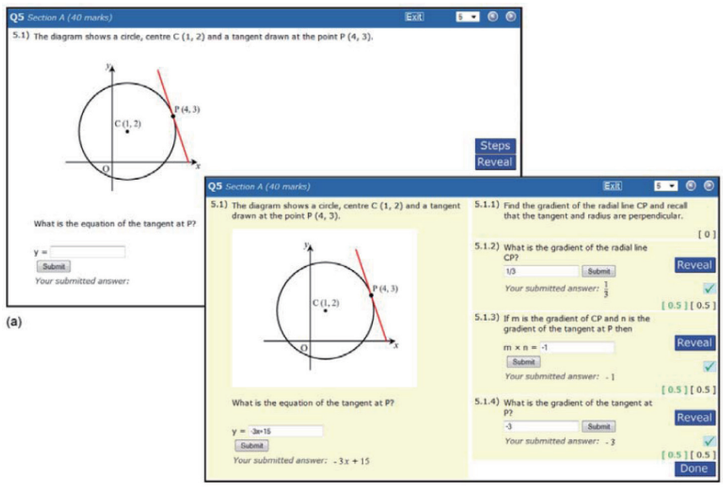

图2（a）显示的是学生在没有任何帮助步骤的情况下完成的一道题；（b）显示的是题目被分解成若干步骤[①]

Ashton et al （2004）发布了学生在使用PASS-IT考试系统的表现的总结报告。对学生来说，看到自己的表现情况能够帮助培养独立学习的能力。这种分析也适合循环问题的设计过程。例如，如果学生在某一特定问题中持续使用“分解步骤”，这就意味着学生对解决该问题是有困难的，也许是因为问题设计得不好，也许是由于学生的误解所造成的。

PASS-IT考试系统的研究和软件的开发现在已经完全融入到了SCHOLAR项目中。在苏格兰的所有中学里，SCHOLAR项目（2013）在中级、中高级和高级学习资料中将形成性考核作为核心。

6. OpenMark 和Moodle: 侧重反馈

英开曾通过CD光盘成功向学生提供互动问题，2002年也开始成功地使用在线系统（Jordan 等，2003；Ross 等，2006），继此之后，2005年英开实施了OpenMark系统。英开的学生人数很多，因此，对在线测评的投资是很有意义的。在远程学习中，及时、有针对性地反馈是十分重要的。

图3展示的是一个典型的OpenMark的问题，这是三个截屏，每个都显示了学生回答问题的尝试过程。这个案例中包含的原则是：

（1）强调反馈；

（2）强调交互性（多次尝试能够使学生对收到的反馈迅速回应）；

（3）支持宽度的交互（提供一系列种类的问题，目的是 “充分使用现代多媒体计算机的全部功能创造参与性的考试”）（Butcher 2008）。

另外，OpenMark考试系统的设计使兼职学生能根据自己的时间、以适合自己日常生活的方式来完成考试。这就意味着考试可以在任何一个点被打断，而日后只要联网便可在任何地点再继续完成（Butcher，2006，2008）。

从2002年开始，英开在课程中引入了互动机考(iCMAs)。在2012年8月，大约有60门独立的课程中设置了630 000多个互动机考(Butcher 等，2013)，这其中，四分之一的考试已经成为课程的正式考试办法（比如说，作为终结性考试或者作为入门考试）。在终结性考试里，如果每次尝试不成功，那么就会被扣分。如果每次尝试有一部分是正确的，那么也同样会获得适当的分数和反馈。

英开理学院是第一个使用OpenMark互动机考

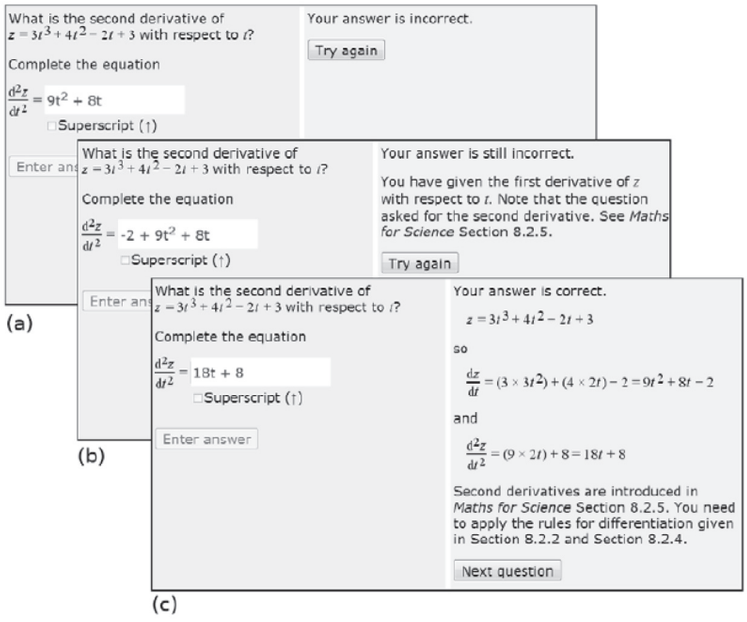

图3 OpenMark考试系统中的一道考题，显示了每次尝试都有反馈。

的部门，因此，也迅速实现了对答案匹配类型题和数字答案准确性的要求（Roass 等， 2003）。机考所提供的恰当的、有针对性的反馈是很重要的。因为只有看到反馈，学生才能够了解到错误的根源。通常来说，学生在第一次尝试之后就会获得反馈，而大部分的反馈能够保留到第二次或第三次的尝试。 Jordan （2011， 2012b）指出改变对问题的反馈有时会影响学生对问题的反应方式。使用统计工具（Jordan 等人 2012）和对个体学生反应的分析（Jordan 2007）能帮助改进问题，并会深入理解学生错误的原因。

OpenMark重点强调提供一系列的问题类型，强调多次尝试、多次反馈，这影响了Moodle考试系统（Butcher 2008）。Moodle的编写样板现在也能够对所有类型的问题提供有针对性的反馈。同时，如果学生在多次尝试以后，Moodle也有能力提供更多的反馈。Hunt（2012）明确了问题种类（例如“数字题”或者“拖拽题”）和问题形式（比如，问题是需要即时反馈、多次尝试的“互动型”，还是只允许一次尝试、直到全部提交才得到反馈的“延迟型”）是互不关联的概念，问题种类和问题形式的结合而产生的测评项是Moodle平台的一个独特特征。

7. 计算机代数系统（包括STACK）

当要测评自由文本中的数学题时，有三种方法来检查学生的答案。最初的CALM考试系统用来评估特定的数值。这是一种合理的方法，但是会导致将错误的答案评为正确的（如：2X和X2在X=2时，结果都是4）。OpenMark采用的是字符串匹配法。这种方法很有效（比如图3所示对学生的答案给出有针对性的反馈），但是却要依赖于题目设置者，他们要考虑到所有正确答案以及尽管结果相同但是答案却是错误的情形。

自从1995年起，一系列的机考系统已经开始采用主流的计算机代数系统（CAS）来检查学生的答案（Sangwin 2013）。例如，AIM采用计算机代数系统Maple（Strickland 2002），CABLE 使用Axiom（Naismith& Sangwin2004），而STACK（使用计算机代数内核教学和测评体系）选择的是开放资源计算机代数系统Maxima（Sangwin 2013）。

STACK作为一个独立的体系于2004年首先发布使用，但是直到2012年该体系才应用到Moodle问题类型中（Butcher 等， 2013）。Moodle系统开发商认为其重要在于对学生答案的反馈和监控方面，这也是STACK体系中重要的内容。在这种情景下，STACK采用了“潜在答案树”，对常见问题能够给出具体的、有针对性的反馈。芬兰阿尔托大学的焦点小组认为及时的反馈是STACK最好的特征（Sangwin 2013），Sangwin对“不是所有使用计算

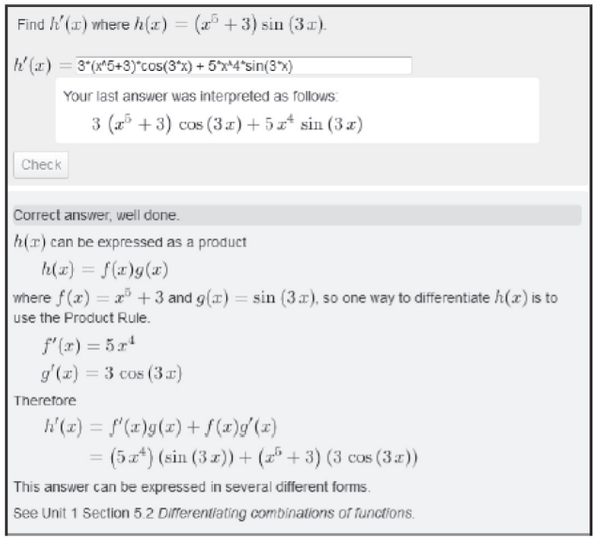

图4 一道回答正确的STACK 问题

机代数系统的（CAS）体系都能够让教师对反馈进行编码”感到非常惊讶。图4阐释了Sangwin (2013)所具体描述的STACK的部分复杂特征。

问题的变式。因为所有的计算都是通过计算机代数系统来执行的，因此，只要稍微努力就可以创作出一个问题的多种不同变式。所以，在上面显示的例子中，变式可能要求学生区分其他两个函数合起来的结果而生成的任何一个函数。然而，最好的做法是检验并使用难度相同的变式。

问题设置者的选择。问题设置者可以决定，例如，是否接受隐式乘法，是插入圆点还是乘号来表示乘法。设置者也可以选择是否接受所有代数方法上相等的答案。所有前述内容如图4所示，但是，要回答“简化U3U5”这个问题，那么答案“U3U5”就是不可接受的。

审核与考试是分开的。学生通过键盘输入的答案会通过系统“理解”而得到显示，在评分之前会检查句法的正确性。而这一过程又使得有些答案有机会欺骗计算机代数系统，在这种情况下，包含“Diff”命令（该命令告诉计算机代数系统区分原始功能）的答案就会被拒绝。

8. 简答题和小论文

在使用计算机代数系统测试更加复杂的数学题的同时，简答题评分软件也被引入到建构题中。“简答”通常是指需要一两个句子来回答，后面紧跟着测评。Jordan（2012b）严格地限制了简答题不超过20个字，一方面告诉学生需要回答什么，另一方面不鼓励既包含正确和非正确答案内容的同时出现。Mitchell et al（2002）最先明确了正确答案中包含的不正确因素是简答题自动评分系统中可能存在的一个非常严重的问题。

简答题评分的软件包括C-rater（Leacock & Chodorow, 2003），智能测评技术（IAT）开发的系统（Mitchell 等， 2002；Jordan & Mitchell， 2009）以及由Sukkarieh、Pulman 和Raikes（2003，2004）所开发的系统。Saddiqi& Harrison（2008）回顾这些系统，在某种程度上来讲，这些系统都是建立在计算机语言基础上的。例如，智能测评技术软件利用信息抽取的自然语言处理（NIP）技术，并依据标准答案的动词和主语模板来比较学生的答案。然而，IAT提供了一种创作工具，即使没有自然语言处理技术的问题，设计者也可以使用。相比而言，OpenMark的PMatch（Butcher & Jordan 2010, Jordan 2012a）和Moodle Pattern Match的问题类型是比较简单的模式匹配系统，基于关键词及其同义词的匹配，有时这种匹配有特定的顺序，可能被一定数量的其他词所分隔，并关注否定词是否出现（如图5所示）。以词典为基础的拼写检查程序告知学生答案中是否包含了不可识别的词，但是允许丢词或者字母顺序调换的标准字符串匹配法仍然有用，以此来解决学生偶尔使用了与预期词汇稍有不同的词汇（例如，使用decease，而不是decrease）。对大多数系统来说，在开发答案匹配中使用真正的学生回答是非常重要的。

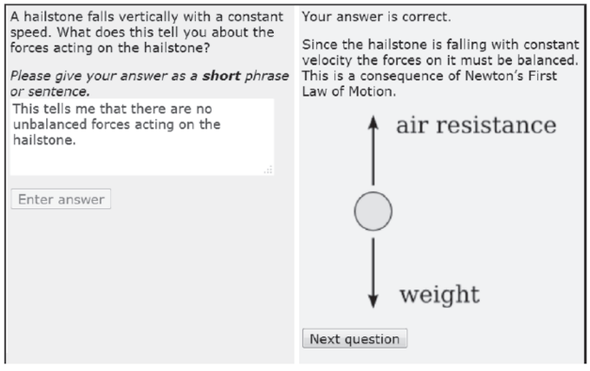

图5 一道PMatch问题的正确答案

英国开放大学OpenMark系统中使用了智能测评技术和PMatch答案匹配法，而PMatch（模式匹配法）与STACK一样，都是Moodle问题类型之一，即时的、定制式的反馈在此类考试中非常重要。关于学生对简答自由文本问题的参与度和所提供的反馈，Jordan（2012b）曾开展过非常详细的评估。

机考可以获得准确评分，好于或至少等同于人工评分（Butcher & Jordan 2010, Jordan 2012a），但此类问题尚未被充分利用。Jordan(2012a)收集了几百个学生答案和评分的需求，认识到此类问题被广泛应用的最大障碍是对设计答案匹配所需的时间。她建议研究应该聚焦于使用机器学习、设计答案匹配规则上；聚焦于调查不同大学生对此类问题给出相同答案的概率；如果他们的答案相似，那么，就具有分享问题的可能性。

小论文自动评分系统与简答题评分系统有显著的不同，前者侧重于写作风格，要求的内容也不如简答题严格。许多系统中都有小论文自动评分功能，如E-rater（Attali & Burstein， 2006）和智能论文评分器（IEA）（Landauer ，2003），Valenti 等 (2003)、Dikli(2006)和Vojak 等 (2011)都曾对这

些系统做过综述。更为智能的系统尚在开发之中，如OpenEssayist（Van Labeke ，2013）侧重于提供反馈帮助学生提升论文写作技能。写作样式采用简单代理模式的系统一直以来饱受批评，如Perelman（2008）就曾经在培训3个学生写小论文时利用这种简单代理模式的窍门，使用长词、引用名句（不管多么的不相关）从而获得机考的高分。Condon（2013）认为直到计算机能够对写作样式做出有意义的评价后，才能使用这种考试方式。

9. 有效使用问题

根据Hunt（2012）描述，机考系统包含三部分：

（1）问题引擎，将每个问题呈现给学生，给学生的答案评分并给出恰当的反馈；

（2）试题库；

（3）测验系统，将单独的问题与完整的测验相结合（有可能在测验阶段给出反馈）。

因此，除了考虑问题类型外，考虑问题的组合方式也很必要。

从题库中选择问题，或者使用每个问题的多种变式能够为考试实践提供更多的机会（Sangwin ，2013），并避免抄袭（Jordan 2011）。然而，特别是在终结性考试中，选择能够评估同一个学习结果并且难易度相同的问题是很有必要的（Dermo ，2010；Jordan 等， 2012）。Dermo(2009)发现：学生随机选题存在一种隐忧。可以在小测验或考试阶段将反馈提供给学生。例如，Jordan（2011）描述了诊断性测试的方法，在该测试中使用了交通信号灯系统表示学生对各种不同技能的准备程度，“红灯”意味着“你还没有准备好”，“绿灯”是指“你好像已经具备了必须的技能”，“黄灯”是指“多保重”。

适性测验（经常被描述为“计算机适性测验”）是根据学生对以前问题的答案来判断其学习能力，从而使接下来所设计的问题能够保持在恰当的水平上（Crisp 2007）。Lilley等(2004)发现学生在适性测验中并没有处于不利地位，而且他们乐意不回答那些自认为简单的问题。计算机适性测验的问题通常是从试题库中选取，使用统计工具来分配难易度（Gershon 2005），因此大部分系统变得很复杂并且依赖大量的、具有标准化试题的试题库。Pyper & Lilley(2010)描述了一个比较简单的“flexilevel”系统，该系统采用固定的分支技术来选择下一个问题以及在哪个阶段呈现给学生。

适性测验的另外一个用法是创造“迷宫”，在此类测验中，后面问题依据学生对前面问题的回答，没有必要给出“正确”或其他答案。Wyllie & Waights（2010）设计了临床决定迷宫，依据各种信息资源，模仿要采取的决定，从而确定如何治疗一位患有腿部溃疡的年老病人。这种迷宫的类型为提高机考真实性提供了一种方法。

海里特一瓦特大学计算机辅助数学学习项目团队成员正在寻找办法，希望在考试中加入模拟元素（Ashton & Thomas， 2006），例如，采用分屏模式，在一边的屏幕上加入实践练习，另一边加上问题（Thomas & Milligan， 2003），分屏的目的是要提高互动性，“在学习环境中学习”来测验学生（Ashton 等， 2006b， p.125）。

另一个将教学与考试相匹配的努力尝试是为教科书匹配题库，最著名的一款此类产品就是培生集团的“精通”系列，如“精通物理”（Pearson ，2013）。尽管此类问题通常被描述为“家庭作业”，但是老师们对如何使用这些问题保留了自己的选择权。培生系统中所提供的对学生回答的分析工具也在学生给出错误信息时起到了引导作用。例如，Walet & Birch（2012）使用了“及时性”教学模式，学生首先通过讲座进行自我学习，然后参与在线测试，测试结果将在几小时后知晓。Walet & Birch使用精通物理，但是发现学生对“家庭作业模式”下的问题（没有提示）并不能很好地理解，所以现在他们使用“辅导模式”（带有提示）并在必要的地方加上反馈。

10. 超越测验的在线测评

技术能够以多种方式来支持测评和提供反馈，在线提交作业、在线人工评分并在线返回给学生。Hepplestone 等人（2011）回顾了该领域的工作，并描述了向学生发布反馈的系统，但是该系统直到学生对所收到的反馈进行反思之后，才会将分数发布给学生。这个办法大大提升了学生对反馈的参与度（Parkin 等， 2012）。同样，也可以使用音频反馈系统（Lunt & Curran， 2010；McGarvey & Haxton，2011）和屏幕转播（Haxton & McGarvey ，2011；O’Malley，2011）来报告学生较好的成绩。

在第三和第四部分讨论了使用PeerWise系统让学生创作和复习问题，但是通常来说同伴测评是指由同伴来评价学生的作业。这种方法可以节省教师

资源，无需教师再给学生提供额外的反馈，但是Honeychurch 等（2013）指出同伴测评的真正价值“不在于反馈本身（反馈作为一种产品），而是进行反馈的过程”。技术能够支持同伴测评，更适合大规模的班级和在线学习环境（Luxton –Reilly，2009；Honeychurch 等，2013）。

电子文件夹、博客、维基和论坛等技术能够用来鼓励学生参与、协作和反思（Bennett 等，2012）。电子文件夹，例如被广泛使用的Pebblepad（2013），能够实现记录跨学科学习过程并反思个人的学习过程，存储成绩。这使得难以测评的内容得到测评，这也鼓励反思性学习方法，让学生对自己的学习负责，并侧重学习中积极的方面（Jafari&Kaufman，2006；Madden，2007）。为了达到测评和反馈的目的，教师们能够查阅学生的电子文件夹，Molyneaux 等 (2009)介绍了一个项目，其中电子文件夹就由一组学生联合管理。文件夹通常与提高学生的就业率紧密相关（Halstead & Sutherland，2006）。

Sim&Hew（2010）确信博客是电子文件夹很自然的一部分，学生通过博客可以分享学习经验并进行反思。Chruchill（2009）鼓励学生参与到博客活动中，并将其作为课程评估的一部分。博客活动受到了学生的欢迎，并有超过一半的学生表示即使不作为课程评估的一部分，他们也会继续自己的博客活动。

Caple & Bogle （2013）热衷于使用维基来评价小组协作活动。小组中的任何一名成员都可以修改维基页面，其优势是每次修改都会被记录下来并归属于做修改的特定使用者。这意味着不但可以评估小组活动，也可以评估个人活动。网上论坛也是很有用的协作工具，但是Conole & Warburton（2005）明确了论坛中“估量”不同互动的难度。

四、在线测评的前景

基于以上对在线测评文献及相关教育技术发展的回顾，我们来预测在线测评的未来发展并讨论可能存在的困难。

1. 大规模开放在线课程(MOOCs)

2013年夏天，Cathy Sandeen在《测评研究与实践期刊》上撰写了MOOCs专题，吸引了我们的注意力，并由此在教育领域掀起了前所未见的热情、实验、讨论和辩论（Sandeen 2013, p.11）。无论MOOCs的未来是什么，积极的一面是此类课程能够促使设计者考虑采用恰当的测评方法，测评大规模的学生非正式和在线学习。

MOOCs最初提供给学生时是免费的并且无学分。在这个阶段，Masters’（2011）指出，“在MOOCs课程中，测评并不能激发学习；学习者自己的目标才是学习的驱动力”，这一论断是合理的。然而，大部分的MOOCs课程都会基于一定程度的参与或者取得的成绩签发某种程度的“徽章”。如果要测量所获得的成绩，那么就需要某种程度的测评。另外，形成性考核提供了参与的可能性并激发学生积极参与，所有的这些因素都可能提高大部分MOOCs课程的完成率。

学习MOOCs课程可以很容易地通过完全在线的机考、协作式和同伴测评方式来运作。机考可以低成本、快捷地传输，但存在低质量测评的风险。为避免这一风险，MOOC体系需要提供多种类的问题，提供有意义的、即时的反馈，允许学生多次尝试问题。另外一个重要的因素是，不同的学生最好接收到不同的问题，这样就需要题库或者问题的多变式。最后，在提供了高质量的工具后，需要保证MOOC创作者经过训练设计出高质量的问题。教师们不应成为“被忽略的学习者”（Sangwin & Grove，2006）。

2. 学习分析和测评分析

学习分析可以被定义为“对学习者及其所处学习环境的数据的测量、收集、分析和报告，目的是理解学习及其发生的环境并对其最优化”（Ferguson， 2012， p. 305），Redecker 等（2012）建议我们应该“超越测验的范畴”，在测评中使用学习分析。在网络环境中收集到的学生互动数据使得掌握学生的实际互动成为可能，从而不需要加入独立的测评环节。Clow（2012）指出学习分析系统甚至可以在非正式的情景中提供如测评一样的反馈。

广义上来讲，学习分析能够向老师告知学生的学习情况，Ellis（2013）呼吁进行“测评分析”（分析测评数据），指出测评在高等教育中无处不在，而在网络学习环境（尤其是社交媒体）中，缺少学生互动测评。Jordan（2013）阐释过网络学习环境（尤其是社交媒体）中开展测评的潜在可能，她分析了学生在机考中的表现，解释了学生的误解，但是也展示了更多深层次参与的内在驱动力。

通过研究两组学生，她发现了对相同作业上存在不同的完成模式，能够将不同模式与不同学生、不同工作量建立联系。

3. 教学、评估和学习之间界限不断模糊

如果如上文所建议的那样，我们使用学习分析作为测评，使用学习分析来帮助学生学习，那么，测评与学习之间的界限就越来越模糊了。同样，更加复杂专业的系统应该应用适性测验，为个性化的形成性考核和诊断性考核（也许也包括终结性考核）提供机会，从而能够更好地支持学生们的学习。

也许，使用移动设备的网上教学能够将教学资源嵌入测评之中，反之亦然。在教学最恰当的时候提出问题，无论何时何地，学生都可以进行测试。考试形式的作业再也不是一个独立的个体。课本内容呈现过程中设置问题是一种常见做法，但是现在这种问题实现了互动。学生可以学习到一系列的教学资源从而帮助他们回答问题，而任何“标准答案”都隐藏起来，直到学生提交了自己的答案才会显示。

4. 计算机的恰当使用

Phil Butcher 在2013年的eSTEeM（2013）年度论坛上回顾了英开机考的使用，“现在是什么样呢？我是否可以建议要开始将计算机当成一个计算机来使用？”他指的具体问题是在Moodle平台中引入了STACK问题类型（Butcher 等，2013）。然而，还有一些近期案例说明计算机能够精准地评分，例如Coderunner（Lobb 2013）通过运行学生所写的编码来测评计算机编程技能。除了用计算机来计算和评估测试本身外，它有可能会利用技术提高问题的质量，例如，从学生答案到自由文本简答问题，采用机器学习来设计答案匹配规则（Jordan 2012a）。

然而，计算机仅当其适用时才能被采用。有时，混合的方法更加有效，数学学科个人学习过程中半自动分析能够自动地监控学生的互动，如果有必要，会将这些互动发送给辅导教师，由其填写详细的反馈（Herding & Schroeder ，2012）。Butcher & Jordan （2010）建议计算机无法“识别的”简答题应该进行人工评分。当前，还有一些测评任务（例如实验报告、小论文、论证）对机评来说仍具挑战。采用小论文评分软件做形成性考核是合理的，但是问题是能否广泛应用于终结性考核，就如McGuire 等（2002）十年前认知的那样，我们不应该错误地认为将问题分解为多个步骤就测评了问题解决过程中的所有技能。概括起来，我们应该利用计算机做其擅长的、解放辅导教师人工评分的部分繁琐劳动，从而使他们去测评那些必须由他们本人去判断真实性的部分。

【注释】

①两幅图是得到海里特-瓦特大学SCHOLAR项目允许后重新设计的图形。CUE是一种考试体系，名称来源于CALM，UCLES （剑桥大学当地联合考试)和EQL (位于苏格兰的一家考试公司）。

【参考文献】

（英文参考文献限于篇幅从略，有需要者详见http://oro.open.ac.uk/38536/）

致谢：

作者非常感谢Tom Mitchell测评技术有限公司的启发和协助，感谢伯明翰大学的Chris Sangwin（STACK系统的开发人），海里特一瓦特大学的荣誉教授Cliff Beevers、教育开发（测评）人Oormi Khaled，尤其感谢来自英国开放大学的Phil Butcher和Tim Hunt。

作者简介：

萨莉·乔丹（Sally Jordan）博士，英国开放大学科学专业高级讲师、责任教师。现任英国开放大学科学院物理学系主任，负责东英格兰地区的各种科学课程的运营。

译者简介：

侯松岩，国家开放大学，教育研究院，硕士。

***原文（E-assessment：Past，present and future）刊于2013年10月《新方向（New Directions）》杂志，，翻译出版时经作者授权，特此感谢。**